PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES
Volume 26 October 15, 1940 Number 10
Copyright 1940 by the National Academy of Sciences

SOME IN VIVO EFFECTS OF LOCALIZED NUCLEAR DISINTEGRATION PRODUCTS ON A TRANSPLANTABLE MOUSE SARCOMA

By Paul A. Zahl,[1] Franklin S. Cooper[2] and John R. Dunning[3]

Communicated June 26, 1940

Introduction.—The destruction of living tissues by x-ray and fast neutron radiation is well known. The effects are directly traceable to the action of energetic electrons resulting from the absorption of the x-rays in the one case, and in the other, from recoil nuclei, such as those of hydrogen, carbon and oxygen which have been projected by neutral impact. However, in both cases the destructive action occurs throughout the irradiated tissue, and no satisfactory method has been found for localizing the damage, in the case of cancer therapy, to the tumor zone. Thus skin damage usually sets an upper limit to the dose which can be delivered through the skin to underlying tissue.

Since the passage of slow neutrons through body tissues is not accompanied by the production of energetic recoil protons, there should be little or no resulting damage to the tissue. However, if these slow neutrons be introduced into a zone which has been perfused with certain chemical elements such as boron or lithium, or their compounds, nuclear capture reactions will occur which release very energetic particles, and result in the local destruction of tissue. That is, for boron:

$$_5B^{10} + {}_0n^1 \rightarrow {}_3Li^7 + {}_2He^4 + 2.2 \text{ MEV (million electron volts)}$$

or for lithium:

$$_3Li^6 + {}_0n^1 \rightarrow {}_2He^4 + {}_1H^3 + 4.6 \text{ MEV.}$$

The energies released by these reactions are approximately 2.2 MEV and 4.6 MEV, respectively. The cross-section[4] for the boron process is 57.5×10^{-24} cm.^{2}, as compared with 60×10^{-24} cm.^{2} for the lithium process. Thus while the energy release for lithium is larger, approximately five times the atomic concentration of lithium is necessary to obtain the same total energy release as with boron. Both reactions are competitive with capture

of the slow neutrons by hydrogen, with emission of high energy gamma rays,[5] ${}_1H^1 + {}_0n^1 \rightarrow {}_1H^2 + \gamma$ (2.2 MEV), but the capture probability for this process is only 0.3×10^{-24} cm.^{2}, so only very small atomic concentrations of boron or lithium are required to capture an appreciable fraction of the neutrons.

The foregoing considerations suggest an investigation of the applicability of neutron-boron or neutron-lithium techniques to the localized treatment of tumors. The physical principles have been discussed at length by Kruger (1) in connection with his experimental findings that mammary carcinoma, lymphoma and an undifferentiated sarcoma showed decreased "takes" subsequent to immersion in boric acid solution and *in vitro* irradiation by slow neutrons.

In an effort to test the effectiveness of the neutron-boron process *in vivo*, experiments were designed to attempt to localize boron or lithium atoms within or around the tumor mass of Mouse Sarcoma 180 during the period of irradiation with slow neutrons. This tumor type was considered suitable for such work because of its known radio-sensitivity and its standardizable characteristics.

Experiments.—The implantation technique together with the growth characteristics of Mouse Sarcoma 180 have been fully described by Sugiura (2), *et al.* Suffice it to say that when inoculated subcutaneously in the axillary region with a 2-mm. cube of freshly excised tumor tissue, male mice of 20–22 grams will ordinarily die during the third or fourth week following implantation. The implanted fragment undergoes rapid growth; metastases are not formed. Death presumably is due to impoverishment of the animal by the growing tissue mass, together with a toxemia syndrome resulting from the by-products of necrosis.

The aim was to inject the tumor with various boron or lithium preparations, followed by neutron irradiation of the whole animal with a dose somewhat below that which would ordinarily kill the animal due to general irradiation effects. It was hoped that by taking advantage of the neutron-boron reaction at the site of the tumor, one could develop a local ionization sufficiently intense to destroy the malignant tissue.

The mouse to be irradiated was placed in a small, perforated aluminum shell. Each shell was placed in a paraffin well capped with a paraffin block. For dimensions of these structures see figure 4. The blocks, piled in stacks of three, were placed within the radiation zone of the cyclotron target, in most experiments about forty centimeters from the target.

The neutrons were produced in the cyclotron at Columbia University by bombarding a beryllium target with protons of approximately 7 MEV energy. The neutrons emitted by this reaction have a spectrum rich in neutrons of 0.5 to 2 MEV and comparatively few have energies above 3 MEV.

Many of the neutrons from the cyclotron were slowed down by impacts with the hydrogen nuclei in the paraffin blocks. The thickness of the paraffin walls was a compromise between that necessary to obtain the maximum slow neutron radiation density in the region of the organism to be irradiated and the minimum ionization due to fast neutrons, taking into account the low intensity of the primary neutron beam. One inch of brass (the chamber wall) reduced the gamma radiation from the target, so that the background ionization due to gamma rays in the region of the tumor was less than 20% of the total ionization under these conditions. The increased ionization in a region containing boron is illustrated in figures 1 and 2. Sufficient space was not available for additional lead, but under more favorable conditions, its proper use could reduce the gamma ionization to a still lower value.

Dosages were adjusted from empirical biological observations, aided by existing information on the reaction of mice to known dosages of x-rays. This was necessary, first, because no accurate measure of the energy release of the slow neutron-boron reaction could be made at the tumor site; second, because the biological effectiveness of this energy is not quantitatively understood.

Before undertaking the boron experiments *per se* it was necessary to determine the lethal limit of irradiation in terms of time for the whole animal. Data given by Lawrence, Aebersold and Lawrence (3) indicate that whole-animal x-ray irradiation of 500 roentgens reduces the average life of mice following irradiation to 17 days; 600 r to 12 days; 700 r to 10.5 days; 800 r to 7 days; 1000 r to 5 days. Because of the scatter around each of these mean periods it was arbitrarily assumed that any of our irradiated animals surviving a period of 35 days had escaped the lethal effects of irradiation, and had not been subjected to a total ionization of more than the biological *equivalent* for the mouse of 400 r of x-rays. The validity of this assumption is borne out by Sugiura's (2) findings that 81% of mice receiving 400 r survive indefinitely.

Experiment 1 of table 1 indicates that 18 hours of irradiation in the paraffin wells was 100% lethal. In experiment 2 the irradiation period was reduced to ten hours, and it was found that 66.6% of the animals survived, indicating an approach downward toward the non-killing threshold. From a study of the Lawrence, Aebersold and Lawrence data and that of Sugiura, it was assumed that during somewhat under a ten hour period our animals were being subjected to an effective biological ionization equivalent of between 300 r and 500 r of x-radiation, which according to Sugiura's data is on the borderline between lethal and sub-lethal. The neutron dosage (without the paraffin filter) as read on a standard Victoreen r-meter was approximately 30 "n" per hour.

TABLE 1

TABLE PRESENTING DATA ON MICE SUBJECTED TO VARIOUS TYPES OF TUMOR INJECTIONS TOGETHER WITH SLOW-NEUTRON IRRADIATION

EXPER. NO.	MATERIAL INJECTED	QUANTITY	RADIATION PERIOD	NO. OF ANIMALS		NO. OF ANIMALS ALIVE AT END OF 35 DAYS	% OF TOTAL LIVING AND SHOWING TUMOR CURE OR REGRESSION	% OF TOTAL DYING FROM TUMOR, GENERAL RADIATION EFFECTS OR NATURAL CAUSES	% CURE OR REGRESSION DUE TO BORON-NEUTRON PROCESS
1	No tumor	...	6 hrs. on each of 3 successive days. Total—18 hrs.	Irradiated	18	none*	..	100.0	..
				Non-irradiated	24	21	..	12.5	..
2	No tumor	...	5 hrs. on each of 2 successive days. Total—10 hrs.	Irradiated	18	12	..	33.3	..
				Non-irradiated	24	22	..	8.4	..
3	Particulate boron in sesame oil	0.1 cc.	5 and 6 hrs. on each of two successive days. Total—11 hrs.	Irradiated	12	2	16.6	83.4	8.3
				Non-irradiated	12	1	8.3	91.7	..
3a	None	...	Same as in No. 3	Irradiated	12	1	8.3	91.7	..
				Non-irradiated	12	2	16.6	83.4	..
4	Lithium metaborate in sesame oil	0.1 cc. 50 mg.	6 hrs.	Irradiated	24	13	54.1	45.9	16.0
				Non-irradiated	21	8	38.1	61.9	..
5	Same as in No. 4	0.05 cc. 0.05 cc. Tot. 50 mg.	3 hrs. on each of two successive days. Total—6 hrs.	Irradiated	18	12	66.6	33.4	22.2
				Non-irradiated	18	8	44.4	55.6	..
6	Boric acid in sesame oil	0.05 cc. 0.05 cc. Tot. 50 mg.	4.5 and 3.6 hrs. on two successive days. Tot. 8.2 hrs.	Irradiated	24	12	50.0	50.0	45.0
				Non-irradiated	20	1	5.0	95.0	..
7	Same as in No. 6	0.1 cc.	7 hrs.	Irradiated	18	10	55.5	44.5	15.5
				Non-irradiated	20	8	40.0	60.0	..

* All animals in this experiment were dead within the first 25 days, presumably from general irradiation effects.

Since the appropriate time-dosage period was ascertained grossly to be between six and ten hours, it was necessary to develop a technique for retaining a high concentration of injected material at the tumor site for long periods. Preliminary experiments designed to test the permeability and diffusion characteristics of various forms of boron and lithium indicated: (1) that a saturated aqueous solution of boric acid injected in or around the tumor would not sustain itself in sufficiently high concentration for more than ten or fifteen minutes, as ascertained by making qualitative analyses for boron content of the tissue by the use of the quinalizarine tests of Feigl (4); (2) that finely pulverized metallic boron particles (in the order of 0.5–2.0 μ in dimension) when injected in oil suspension would localize largely at the oil-tissue interface and would not diffuse through the cell membranes or far into the intercellular spaces of the compact tumor tissue. Since the range of the disintegration products of the neutron capture processes was limited to less than fifty microns, the ionizing effect during radiation would be too local for a general tumor-killing effect. (3) That when powdered boric acid suspended in oil in liquid-paste form (one gram doubly pulverized boric acid suspended in two cubic centimeters of sesame oil) and injected into or around the tumor, a large excess of the boric acid could be localized and remain harmless to the tissue, slowly being taken into aqueous solution by the body fluids bathing the tumor and its environs. Thus a relatively high concentration of aqueous boric acid could be sustained at the tumor site for as long as several hours, draining out of the oil suspension into the soluble water phase. (4) That when lithium meta-borate (which is much less water-soluble than boric acid, and which in aqueous solution hydrolyzes into lithium hydroxide and boric acid) is likewise suspended in oil and injected, it goes into body solution even more slowly than the boric acid oil suspension, and therefore was considered suitable for injection preceding very long radiation periods. The toxicity of these materials will be discussed subsequently.

The exact method of injection is a point which warrants description. The needle was inserted under the skin about an inch posterior to the tumor, then passed subcutaneously into the tumor site and through the center of the tumor mass, and extended into the connective tissue anterior to the tumor. The needle was then slowly withdrawn as the plunger was applied so as to deposit material anterior to, within and posterior to the tumor. Before withdrawing the needle from the skin more material was deposited on either side of the tumor. The purpose, of course, was to bring the actively proliferating tumor tissue into close contact with the boron or lithium-bearing oil.

Having found the time-dosage threshold, it was necessary to establish the independent effects on the tumor of each of several variables. The first was radiation itself without injection. Twelve animals with growing, week-old

* All animals in this experiment were dead within the first 25 days, presumably from general irradiation effects.

tumors were irradiated for eleven hours. Only 8.3% of these survived the 35-day period, the others dying presumably because of the sarcoma growth. In the non-irradiated control to this experiment 16.6% of the animals survived. Since the difference between 8.3 and 16.6, in view of the size of the sample, is hardly significant, it can be assumed that the whole body irradiation did not affect a significant regression of the tumors. This is rather what would be expected in view of the fact that from Sugiura's data it requires an equivalent of 1000 r to effect a 50% *in vivo* regression in shielded animals.

The second variable was the toxicity of the injected materials on both the whole animal and the tissue at the site of the injection. Preliminary experiments indicated that amounts of both boric acid and lithium metaborate could be injected subcutaneously into healthy mice in considerable excess to the 25–50 mg. doses injected for the experiments, without any observable ill effects. However, in experiments 4, 5, 6 and 7 it will be seen that the non-irradiated tumor-bearing mice receiving control injections equal to those in the irradiated specimens showed tumor regression or ab-

DESCRIPTION OF PLATE

Figures 1 and 2. Increased ionization produced in regions containing boron when bombarded with neutrons. Fig. 1: Without boron. Eastman Alpha Particle Spectroscopic plate, emulsion No. 129,975. Film exposed to cyclotron irradiation within aluminum shells and paraffin-well in position identical to that occupied by mice in the experiments described in the text. Three tracks due to protons projected by neutrons are clearly visible in the field, together with general photographic grain reduction due to protons projected at various angles to the plane of the film, together with projected nuclei such as carbon, and to gamma-ray background. This photo illustrates essentially the amount of ionizing energy released in an equivalent volume of hydrogen-rich tissue during an equivalent period under conditions described in the text when no boron is present. Magnification: × 1000. Fig. 2: With boron. Same type of film and same exposure time as in figure 1, except emulsion before exposure was dipped into a 2% aqueous solution of boric acid and allowed to dry. Shows numerous alpha particle tracks resulting from slow neutron-boron capture. Many tracks do not lie in plane of the photograph. Illustrates essentially the amount of ionizing energy released in an equivalent volume of tissue in the area of the tumor in which an equivalent concentration of boron following injection with boron salts was maintained. Cf. figure 1. It is the ionization differential illustrated in these two photographs which presents the basic rationale of the experiments described in the text.

Figure 3. Cyclotron with paraffin bricks *in situ*. In each brick are two wells of dimensions indicated in figure 4. Loosely placed in each well is a perforated aluminum shell for housing the living mouse during the radiation. The bricks, in stacks of three, were placed in an arc around the beryllium target in a position designed, so far as possible, to equalize the dosage for each animal. Thus twenty-four mice could be irradiated at one time.

Figure 4. Close view of paraffin brick and wells used in all the experiments. A mouse enclosed in the aluminum shell could be retained in relatively comfortable confinement for more than six hours. The wells were closed from above with paraffin slabs loosely fitted for ventilation.

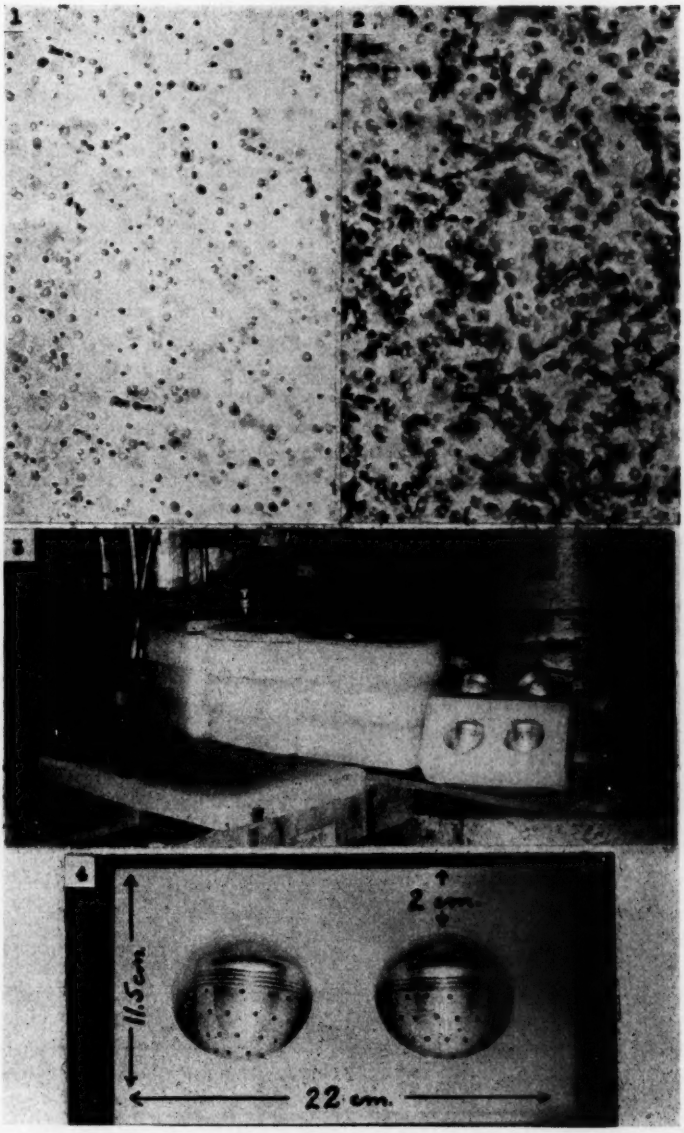

PLATE 1

See opposite page for description of plate.

sorption considerably higher than is normally found in non-injected and non-irradiated tumor-bearing mice. In the four experiments cited the percentages of regressions following injection alone were respectively 38.1, 44.4, 5.0 and 40.0, as compared with the 16% regression in the untreated mice.

This is taken to mean that one or both of the following factors are operative in causing this regression: (1) mechanical injury of the inserted needle to the growing tumor tissue, to the blood channels feeding it or to the connective ground tissue; (2) toxic or destructive effects of the injected materials on the growing tumor tissue, on the blood channels feeding it or on the connective ground tissue. One is inclined to believe that a combination of these two factors is responsible for the increased regression incidence following mere injection of the materials under question.

It is obvious that only percentages of regression significantly above that caused by these two factors can be considered as due to the slow neutron-boron process in the experiments which combine the irradiation and the localization of the boron or lithium compounds within or around the tumor. A study of the data of experiments 4, 5, 6 and 7 of the table clearly indicates that the regression in mice receiving both irradiation and the boron or lithium injections is considerably higher than in those receiving the injections alone, the percentage differences in these four cases being respectively 16.6, 22.2, 45.0 and 15.5. The differences or boron-neutron cures are not large, but since in all four experiments they are consistent, their significance seems conclusive.

Discussion.—The actual variation in the magnitude of these regression differences seems to be subject to any of several explanations: (1) Although the animals, in relation to the source of the irradiation, were placed in such a position so as nearly as possible to equalize the dosage of all the animals being irradiated at one time, complete uniformity was impossible to achieve, due to the architecture of the cyclotron. (2) Fluctuations in the output of the cyclotron during the excessively long radiation periods. (3) Variations in the effectiveness of the injections: the amount of injury, and diffusion factors of the injected materials. (4) Variation in the normal frequency of regression.

The ionization at the site of the injection was not known. As cited above from Sugiura's experiments, a radiation equivalent to 1000 r of x-rays is necessary to cause a 50% regression *in vivo*. With 750 roentgens of x-rays Sugiura reports a 20% regression *in vivo*. Considering the magnitude of our regressions as between 15.5 and 45.0 per cent, and postulating the actual destruction of the malignant cells, it is necessary to assume that an ionization approximately equivalent to 750–1000 roentgens of x-rays was being achieved within or at the tumor, whereas the whole body was being subjected to less than the equivalent of 400 r. We assume that this energy

difference was the result of the ionization caused by the disintegration products of the slow neutron-boron process.

On the other hand, due to the peculiar selective permeability properties of the living cell membrane, we find it difficult to believe, following the injection of boric acid or lithium meta-borate into and around the tumor mass, that as the material slowly goes into aqueous solution a high concentration is maintained uniformly both within and out of the malignant cells. Even if material were injected and retained at the site of the tumor, it was early doubted whether the boric acid or lithium hydroxide would diffuse uniformly throughout the tissue mass in concentrations necessary to be effective in capturing a large proportion of the slow neutrons. Indeed, the question of whether the ions would easily permeate the membranes of tumor cells has not yet been established, and there is little evidence to indicate that they would concentrate in sufficiently large quantities to be effective. It is more likely that they diffuse into and through the intercellular spaces. But in the case of a compact tumor such as Mouse Sarcoma 180 the amount of intercellular diffusion is limited.

For this reason, and because of the relatively high incidence of ulceration around the tumor following the injection-radiation treatment, at the present time the curative regression is interpreted as due to one or a combination of the following factors: (1) alpha particle or proton destruction of some or all of the malignant cells, (2) impairment of the vascular system feeding and draining the tumor, (3) radiation effect on the connective tissue base of the tumor.[6]

One of the interesting aspects of this work is that, unlike x-ray therapy where shielding of tissues not under treatment must be carefully applied, we were able to subject the whole body to the same extrinsic energy as the tumor; but that because of the neutron-boron process we were able to set up a high ionization differential between the tumor and the rest of the body.

The authors consider that for any possible future employment of the slow neutron-boron process in tumor therapy, some device other than simple hypodermic injection should be developed for localizing either boron or lithium or related materials in malignant tissue. This is particularly essential in the case of involved metastasizing and deeply situated growths. Experiments involving the intravenous injection of large particle colloidal dyes to which the lithium atom is attached have indicated considerable localization of lithium in spontaneous mouse tumors. This work is being continued.

It is to be mentioned also that concurrently with the work described in this communication, experiments were undertaken in which boron in one form or another was injected into the mouse testis, followed by slow neutron irradiation. A clearly observable effect on the germinal cells of the seminiferous tubules much more extensive than either the radiation effect alone

or the effect of the chemicals also was observed. These results will be published elsewhere.

Summary.—Transplantable mouse sarcomas were injected with various forms of slow-neutron-capturing materials. When the whole animal whose tumor was so injected was irradiated with slow neutrons, a significant increase in tumor regression was observed. This increase is attributed to the localized ionization resulting from the nuclear disintegration products of the capture process.

LITERATURE CITED

(1) Kruger, P. Gerald, "Some Biological Effects of Nuclear Disintegration Products on Neoplastic Tissue," *Proc. Nat. Acad. Sci.*, **26**, 181–192 (1940).

(2) Sugiura, K., "The Effect of Roentgen Rays on the Growth of Mouse Sarcoma 180 Irradiated *in Vivo*," *Radiology*, **28**, 162–171 (1937).

(3) Lawrence, J. H., P. C. Aebersold and E. O. Lawrence, "Comparative Effects of X-Rays and Neutrons on Normal and Tumor Tissue," *Proc. Nat. Acad. Sci.*, **22**, 543–557 (1936).

(4) Feigl, Fritz, *Qualitative Analysis by Spot Tests*, Nordemann Publishing Co., New York (1939).

(5) Sugiura, K., "Studies on Radiosensitivity of Mouse Sarcoma 180 Irradiated *in Vivo* and *in Vitro*," *Radiology*, **29**, 352–361 (1937).

[1] Memorial Hospital, New York City. Work supported by The Haskins Laboratories.

[2] The Haskins Laboratories, New York City.

[3] Columbia University. Aid of the Research Corporation is gratefully acknowledged. We wish to express our appreciation to Dr. C. P. Rhoads and Dr. G. Failla for courtesies extended to us at Memorial Hospital, New York City.

[4] The cross-section is a measure of the probability of interaction of the neutron with the nucleus, i.e., if there is a flux of one neutron per centimeter squared per second and one atom per cubic centimeter, the cross-section represents the probability of this neutron being captured by the atom.

[5] Other elements in tissue also capture some slow neutrons but the resulting contribution to *local* ionization is not extensive. It should be noted that the small absorption within the tissue of the 2.2 MEV γ-radiation resulting from slow neutron capture by hydrogen largely offsets the efficiency of this capture process, at least as regards local ionization.

[6] Sugiura (5) and others have demonstrated that the incidence of tumor-takes is significantly lower in areas which have been previously subjected to irradiation, thus indicating the importance of the physiological condition of the tissue base to the proper growth of introduced cancerous tissue.

GALACTIC AND EXTRAGALACTIC STUDIES, X. NOTE ON LATITUDE EFFECT AND RADIAL GRADIENTS IN THE NORTHERN GALACTIC HEMISPHERE

By Harlow Shapley and Rebecca B. Jones

Harvard College Observatory

Communicated September 10, 1940

1. The preceding paper[1] of this series summarized the magnitudes of some 22,000 galaxies in a five-degree zone encircling the sky at declination $+43^{\circ}.6$. The material is suitable for an examination of the dependence of nebular density on galactic latitude, and for the study, through the relative frequency of magnitudes, of radial density gradients at different places along the zone. From a recent survey in the south galactic polar cap,[2] the latitude effect, if present at all, was found to be smaller than other transverse gradients in that region. In an earlier study[3] of "sample regions" in both hemispheres, in which low as well as high latitudes were involved, we found, in essential agreement with Mount Wilson results, the well-known avoidance-zone effect up to latitudes $\pm 30°$, but in the northern galactic hemisphere no appreciable latitude effect from $\beta = +30°$ to the galactic pole.

TABLE 1

DISTRIBUTION CONSTANTS FOR 18 PLATES

PLATE	β	b	m_1	PLATE	β	b	m_1
MC 29493	$+35^{\circ}.5$	0.70	15.58	*MC* 30038	$+73.6$	0.77	15.30
29215	39.8	0.57	15.16	30129	70.2	0.69	15.37
27687	46.9	0.56	15.13	29962	65.4	0.73	15.60
30078	52.2	0.62	15.18	29979	60.8	0.53	15.05
29265	58.3	0.70	15.33	28949	56.2	0.48	14.82
29283	63.8	0.68	15.30	30079	49.2	0.64	14.91
28788	68.3	0.76:	15.68:	30186	42.7	0.59	14.85
27639	72.1	0.65	15.03	29537	37.5	0.56:	15.02:
27704	74.0	0.64	15.29	30123	31.6	0.62	15.28

2. In the present discussion earlier results are corroborated. For latitudes greater than $+30°$ the irregularities in distribution from place to place (figures 2 and 3) are sufficient to conceal whatever progressive increase there may be in numbers of galaxies with latitude, such as should result from a uniform or approximately uniform layer of absorbing material along the galactic plane. We find also that for the whole area involved the average number of galaxies at each magnitude down to 17.5 differs but little in high latitudes from the average found in the southern galactic hemisphere (equations (1) and (2)). But the space density increases conspicu-

ously with distance for the section of the zone between 10^h and 14^h in right ascension (table 2).

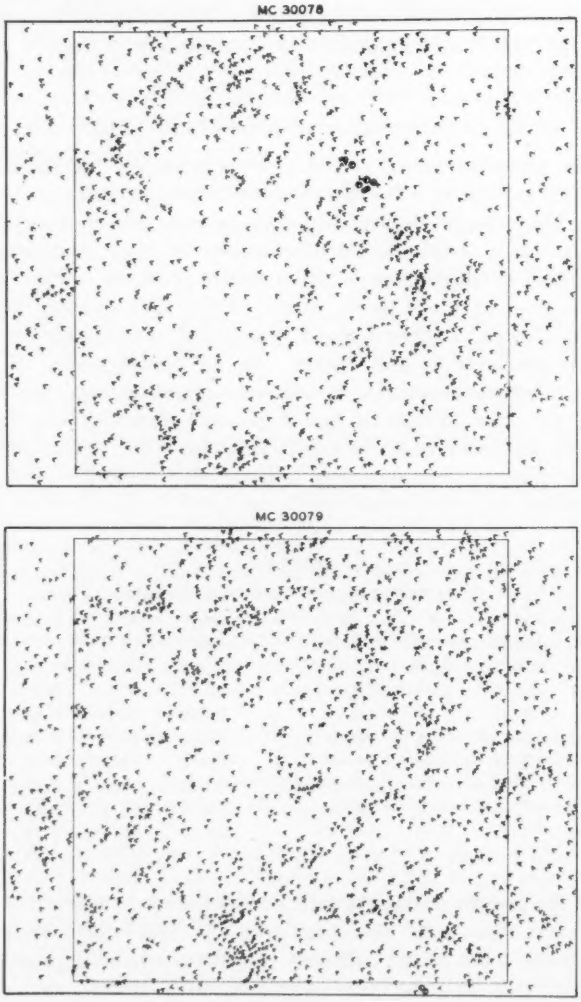

FIGURE 1

Nebular distribution on two plates. Inner areas cover twenty-five square degrees.

The surface distribution is illustrated in figure 1 for two average plates. Stars are not indicated; the positions of NGC objects are circled. The

nebular groupings shown on these plates have not heretofore been recognized as clusters of galaxies.

3. Figure 2 shows the nebular distribution around the zone. For each plate the number of objects per square degree to magnitude 17.5, within the central twenty-five square degrees, is indicated by the ordinates; both right ascension and galactic latitude are used as abscissae. The twenty-five square degrees of each plate is treated as a unit, both in this diagram and in the following discussion. We can see from the figure that within ten degrees of the galactic circle there are practically no galaxies at either of the two crossings of the Milky Way. The relation of nebular density to galac-

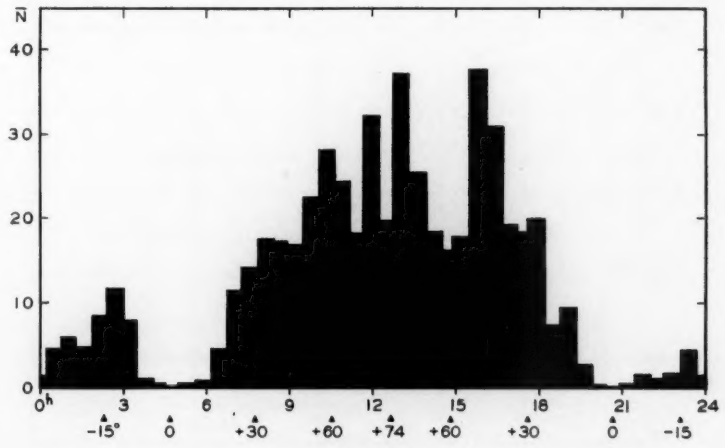

FIGURE 2

Distribution of galaxies to magnitude 17.5 along the five-degree zone. Abscissae are both hours of right ascension and degrees of galactic latitude; ordinates are numbers of galaxies per square degree.

tic latitude is further illustrated in figure 3, where the logarithm of the average number of galaxies per square degree is plotted against galactic latitude for three magnitude limits.

4. If we consider the space absorption and its unevenness as unimportant for latitudes higher than $\pm 30°$, we can use this homogeneous material, after correcting for the red-shift, for a study of the nebular distribution parameters, b and m_1, in the relation $\log \bar{N}_m = b(m - m_1)$. From plots giving for each plate the apparent photographic magnitude, m, against the logarithm of the number of objects equal to or brighter than that magnitude, $\log N_m$, the values of b and m_1 have been determined graphically for the eighteen plates with $\beta > 30°$; they are given in table 1. These values

are derived on the assumption that the surveys are complete to magnitude 17.5; for the four plates (table 1 in the ninth paper[1]) for which the limits of completeness, m_n, are a little brighter than magnitude 17.5, appropriate corrections (through the uniform density assumption) have been made. The values of b, the radial density gradient, vary from 0.48 to 0.77, and in the mean $\bar{b} = 0.64 \pm 0.02$ (m. e.); the values of m_1, the space density parameter, vary from 14.82 to 15.68, with the mean $\bar{m}_1 = 15.22 \pm 0.06$ (m. e.).

If we plot the logarithms of the sums for all eighteen plates (four hundred fifty square degrees) against the apparent magnitudes, we have the result

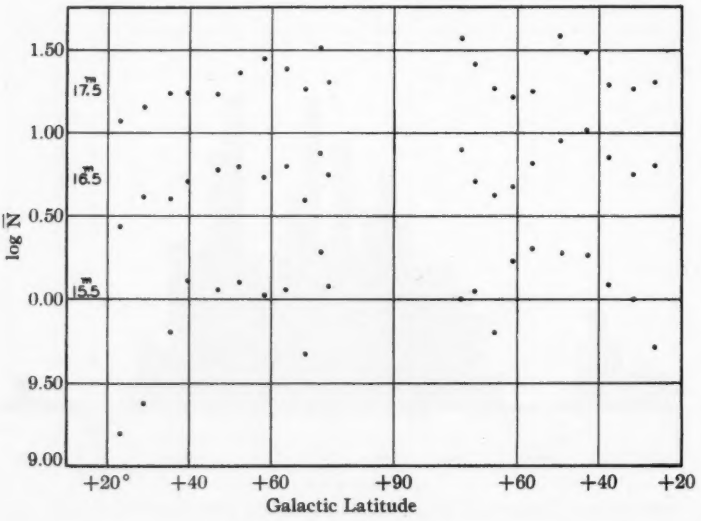

FIGURE 3

Variation of the logarithm of the number of galaxies per square degree with galactic latitude for three magnitude limits. All plates with galactic latitude greater than $\pm 20°$ are included.

shown in figure 4. The best straight line representation, graphically determined, gives $b = 0.62$ and $m_1 = 15.20$; a least-squares solution, in which the successive cumulative totals are weighted according to the number of galaxies involved, gives the values that we shall adopt as mean for the zone:

$$\begin{aligned} b &= 0.625 \pm 0.017 \text{(m. e.)} \\ m_1 &= 15.18 \pm 0.04 \end{aligned} \qquad (1)$$

These values should be compared with those obtained for the 198 square

degrees in the south galactic cap,[2] in a survey that extends to somewhat fainter galaxies:

$$\begin{aligned} b &= 0.592 \pm 0.009 \text{(m. e.)} \\ m_1 &= 15.16 \pm 0.02 \end{aligned} \qquad (2)$$

The similarity in the values of m_1 in equations (1) and (2) can be taken either as a measure of the essential equality in density in regions that are separated by something like fifty megaparsecs, with the implication of equality of space absorption in the northern and southern directions; or that greater absorption on one side is compensated by low density on the other. The difference between the values of b, 0.033, is not much larger than its mean error, but the steeper gradient on the north may be significant in the local structure of the metagalaxy.

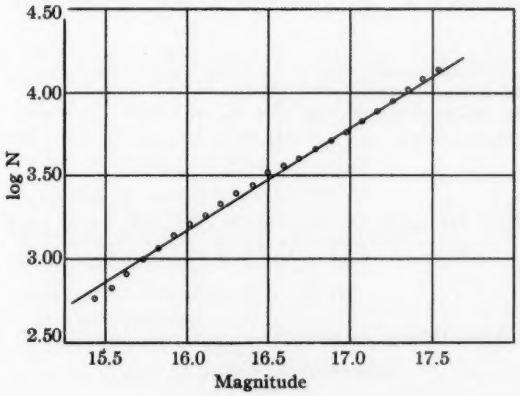

FIGURE 4

Frequency of magnitudes, corrected for red-shift, for the eighteen plates of highest galactic latitude (450 square degrees). Abscissae are photographic magnitudes; ordinates are logarithms of cumulative total numbers of galaxies.

The close similarity of average density at magnitude 17.5 in the two hemispheres may be quite accidental and disappear altogether when larger areas of the high latitude sky are surveyed. In fact, the values of b and m_1 vary considerably from plate to plate along the 5° zone, as shown in table 1. In table 2 we give results of determinations by least squares of the parameters when the material from the eighteen plates is divided into early and late right ascension groups, as well as when the nine plates of highest galactic latitude are treated as a group. Practically the same results would be obtained if the means of the graphically determined values in table 1 were computed for the same groups of plates.

TABLE 2

MEAN DISTRIBUTION CONSTANTS

RIGHT ASCENSION INTERVAL	NUMBER OF PLATES	b	m_1
$8^h15^m.0 - 12^h27^m.6$	9	0.646 ± 0.015(m. e.)	15.27 ± 0.03(m. e.)
13 04.1 – 17 26.6	9	0.627 ± 0.020	15.12 ± 0.04
10 22.1 – 14 42.8	9	0.682 ± 0.010	15.32 ± 0.02
8 15.0 – 17 26.6	18	0.625 ± 0.017	15.18 ± 0.04

For the group of nine plates of highest latitude (225 square degrees containing 5504 galaxies of magnitude 17.5 and brighter) the space density, $m_1 = 15.32$, is the lowest. The gradient, $b = 0.68$, is much steeper than for the other combinations—possibly an effect of unrecognized clusters or clouds of galaxies that begin to appear near the fainter magnitude limit on some of the plates. The full picture of local metagalactic structure in these northern regions must await considerable further work on nebular magnitudes.

[1] These PROCEEDINGS [ninth paper of the series, Sept., 1940].
[2] Op. cit., **26**, 166–176 (1940).
[3] Op. cit., **19**, 389–393 (1933); *Harv. Bull.* 894, 10 (1934).

THE EFFECT OF SHEAR INSTABILITY ON THE TRANSVERSE CIRCULATION IN THE ATMOSPHERE

BY H. J. STEWART

GUGGENHEIM AERONAUTICAL LABORATORY, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated August 20, 1940

The motion of the atmosphere can be considered as a mean flow which is of very large scale and is only slowly changed and, superimposed on this the low level, smaller scale phenomena usually associated with the polar front. If the mean pressures over a period of about a week are plotted, it is seen that the latter disturbances are averaged out, and only the large-scale mean motion is shown. Such a plot of the Northern Hemisphere shows, in addition to the mean westerly flow of air, large-scale closed isobaric systems spaced at comparatively regular distances on the surface of the earth. These include the Aleutian and Icelandic low-pressure areas to the north of the westerlies and the Pacific and Bermuda high-pressure areas to the south of the westerlies. The maximum number of high-pressure cells seems to be about six.

As the position and strength of these disturbances apparently control the mean path of the low-level storms, a knowledge of the properties of these systems would appear extremely desirable for any long-range weather

forecasting. One aspect of this problem has been investigated by Rossby[1] who showed that wave disturbances of about the observed wave-length can exist as a result of the variation of the Coriolis acceleration with latitude. In the present paper, the results of an investigation of another aspect of the problem arising from the existence of the shearing motion on either side of the belt of westerlies will be given. This investigation was started as a result of conversations with Th. von Kármán and C. G. Rossby and was carried out in coöperation with the United States Department of Agriculture under the provisions of the Bankhead-Jones Act.

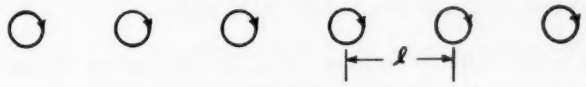

FIGURE 1

In this investigation it was assumed that the atmosphere could be treated as a single layer of fluid of constant density with the vertical velocities being of small importance so that the pressure was determined by the hydrostatic law. In addition the variation of the Coriolis parameter with latitude and the effects of friction were neglected.

The notation used in the discussion is as follows:

x, y = Cartesian coördinates on a rotating disc
r, θ = Polar coördinates on a rotating disc
u = velocity in x direction
v = velocity in y direction
ω = angular velocity of disc
h = depth of fluid
g = acceleration due to gravity
$K_0(x)$ = modified Bessel function of the second kind.

If the motion could have started from rest with a uniform depth, h_0, the principle of conservation of absolute vorticity states that

$$\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{2\omega}{h_0}(h - h_0). \tag{1}$$

If the velocity components are eliminated from this equation by means of the geostrophic wind equations,

$$-2\omega v = -g\frac{\partial h}{\partial x}$$

and (2)

$$2\omega u = -g\frac{\partial h}{\partial y},$$

an expression determining the depth of the atmosphere (i.e., sea-level pressure) is obtained. This is

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} - \frac{4\omega^2}{gh_0}(h - h_0) = 0. \tag{3}$$

The only steady state solution of this equation which vanishes at infinity and which represents flow in circles about the origin is

$$\frac{h - h_0}{h_0} = AK_0\left(\frac{2\omega r}{\sqrt{gh_0}}\right) \tag{4}$$

where A is an arbitrary constant. If A is positive, the motion is anticyclonic; if A is negative, the motion is cyclonic. All the motions considered in this investigation were produced by superposition of vortices of these types.

The shear field south of the westerlies can be considered to a first approximation as a sheet of vorticity separating a moving current from a stationary one. Such a vortex sheet was shown to be dynamically unstable. A similar calculation by Pekeris yielded the same result. This means that the shear fields bordering the westerlies tend to break up into discrete eddies. If there should be stable formations of vortices, dynamic forces would cause these small eddies to collect into one of the stable formations. An investigation of the stable vortex formations should thus give a clue to possible patterns of motion.

For small enough distances from the vortex given in equation (4) or, more exactly, if $\frac{2\omega r}{\sqrt{gh_0}} \ll 1$, the velocity varies inversely with the radius just as in an ordinary two-dimensional vortex in an incompressible fluid without Coriolis forces, and similar results regarding stability will be obtained provided the vortex spacing is such that the above-mentioned inequality is satisfied; however, the geostrophic wind equations, equation (2), are greatly in error in this region so the results of investigations with ordinary vortices cannot be applied directly.

A single infinite row of vorticles as in figure 1 is readily seen to be in equilibrium, but investigation proves the equilibrium to be unstable, and this type of pattern could not be expected to form.

As there is a shear field on either side of the westerlies, a double row of vortices is a better fit to this shear field. There are two possible equilibrium arrangements for such a double-row system as shown in figure 2. Such systems are similar to the Kármán "vortex street" and were investigated in a similar manner.[2] The results of this calculation show that the symmetrical arrangement is unstable while the asymmetrical arrange-

ment is stable for a certain range of values of d/l which depends on the value of $\frac{2\omega l}{\sqrt{gh_0}}$ as shown in the following table:

STABLE RANGE OF VALUES OF d/l FOR VORTEX STREET

$\frac{2\omega l}{\sqrt{gh_0}}$	UPPER LIMIT	LOWER LIMIT
0	0.281	0.281
1	0.300	0.281
2	0.350	0.280

This result is shown graphically in figure 3. It should be noted that for small values of $\frac{2\omega l}{\sqrt{gh_0}}$ the result, $d/l = 0.281$, is exactly that obtained by Kármán. The flow pattern associated with the vortex street shows a jet of westerlies with the vortices drifting to the east.

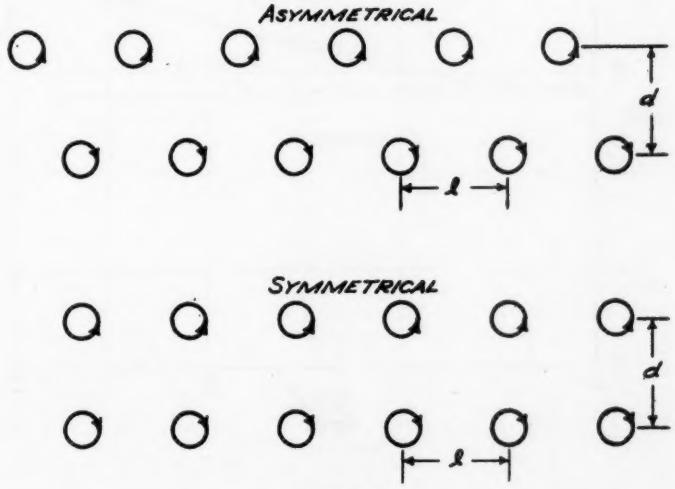

FIGURE 2

These results are directly applicable to the spherical earth only when the distances l and d are small compared with the earth's radius. This is equivalent to saying that the general circulation is very weak.

The observed pressure systems have spacings of the order of the earth's radius, and it is necessary to consider the fact that the path of the westerlies is actually curved to investigate motions on this scale. A general solution of the problem of the stability of double-ring systems is very difficult, but the nature of the results can be estimated by considering some limiting

cases. The preceding calculation gives a limiting case where the circulation is very weak. Normally the shear field to the north of the westerlies is weaker than that to the south. If the northern shear can be entirely neglected, the only possible steady pattern consists of equal anticyclones on the corners of a regular polygon and of a constant distance, a, from the north pole as in figure 4. Investigation shows that such a system is stable for two, three, four, five or six vortices. For seven vortices the arrangement is stable if $(2\omega a/\sqrt{gh_0}) > .71$. For more than seven vortices, the arrangement is unstable. For small values of $2\omega a/\sqrt{gh_0}$ this is in agreement with the work of J. J. Thomson.[3] As values of $2\omega a/\sqrt{gh_0}$

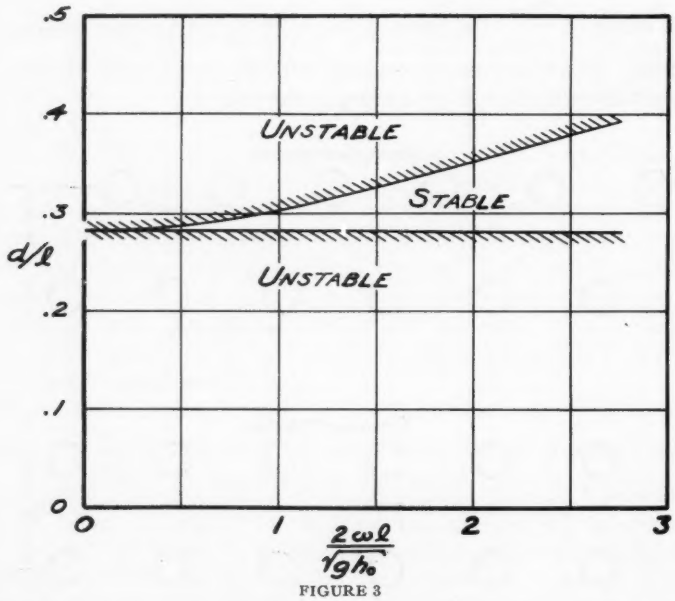

FIGURE 3

greater than .71 are of no apparent significance, these results may be summarized by saying that a ring of six or less vortices is stable. The flow pattern associated with a ring of anticyclones shows a wide jet of westerlies with the anticyclones drifting westward.

The effect of the shear field north of the westerlies may be roughly estimated by placing a single fixed cyclonic vortex at the pole as in figure 5. Such a vortex pattern gives the same results as the single ring provided the cyclone is weak enough; however, as the polar cyclone becomes stronger, the maximum number of anticyclones decreases. If the polar cyclone is strong enough to halt the westward movement of the anti-

cyclones, the system is unstable even with only two anticyclones. This configuration shows a wide band of westerlies and corresponds to a state of strong general circulation. This should correspond quite closely to the observed motion in the atmosphere as the shear field north of the westerlies is weaker than that to the south. As the polar air mass tends to rotate

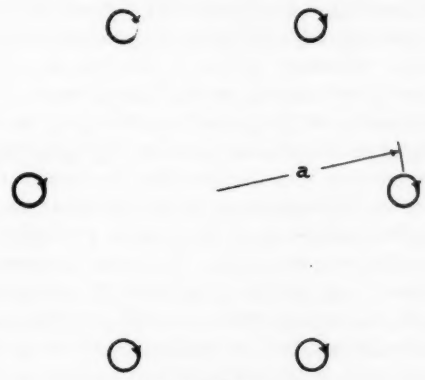

FIGURE 4

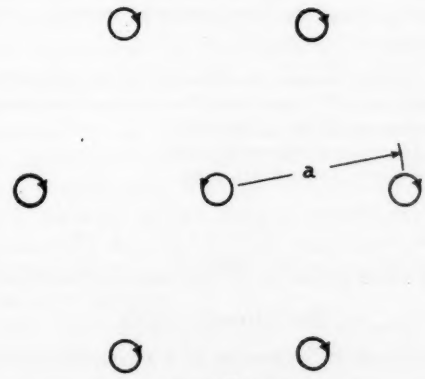

FIGURE 5

as a whole, an increase in the westerly winds, on a percentage basis, increases the shear to the north of the westerlies faster than that to the south. Thus a strong atmospheric circulation corresponds to a large value of the ratio of the vorticity north of the westerlies to the vorticity south of the westerlies, and a weak atmospheric circulation corresponds to a small value of the ratio. From the preceding results, this implies that with

strong circulation only two or three high-pressure cells would be found, whereas with weak circulation stable forms having up to six cells could be found.

As a result of these calculations it is evident that the shear fields next to the westerlies will roll up and form a series of closed isobaric systems with high-pressure cells to the south of the westerlies and low-pressure cells to the north of the westerlies. It appears that a maximum of six high-pressure cells will be observed although more are possible if the belt of westerlies were to become extremely narrow. The low-pressure cells will be asymmetrically placed with respect to the high-pressure cells as in figure 2. Of the two limiting cases, one progressed westward and the other eastward, so one might expect the intermediate systems to be practically stationary. This is in rough agreement with the idea that the position of such systems ought primarily to be determined by thermodynamic considerations, so that the stable vortex configurations found in the atmosphere might be expected to be those that are stationary. It further appears that only two or three high-pressure cells will be found with the strongest atmospheric circulation with the maximum number of cells increasing as the strength of the circulation diminishes, and one might expect a strong inverse correlation between the strength of the circulation and the number of high-pressure cells. This result is well verified by Northern Hemisphere mean-pressure charts although no attempt has been made to correlate these factors in a quantitative fashion.

[1] C. G. Rossby, "Relation Between Variations in the Intensity of the Zonal Circulation of the Atmosphere and the Displacements of the Semi-Permanent Centers of Action," *Jour. Marine Research*, **2**, No. 1, 38 (1939).

[2] W. F. Durand, *Aerodynamic Theory*, **2**, p. 342.

[3] J. J. Thomson, *Motion of Vortex Rings*, p. 94.

ON THE MORTALITY IN HUSBANDS AND WIVES

BY ANTONIO CIOCCO

NATIONAL INSTITUTE OF HEALTH, U. S. PUBLIC HEALTH SERVICE

Communicated September 10, 1940

The study of morbidity and mortality in husbands and wives when properly controlled by other pertinent data offers the possibility of throwing some light regarding the action of a number of factors, environmental as well as hereditary, on the diseases about which so little is known and that are so prevalent in adult and old age. This consideration has served as a stimulus for the initiation of the study reported in this note.*

The material for this study has been obtained from the death records of

1898 to 1938 filed in the Washington County (Maryland) Health Department. The procedure of collecting the necessary data was as follows:

1. Transcripts were made of all the death records concerning white persons stated to have been (*a*) married and (*b*) widowed at the time of death.

2. The records for the married and for the widowed persons, respectively, were alphabetized and divided according to sex.

3. The records of married males have been matched with those of widowed females and the records of widowed males with those of married females.

To date the records of 2578 couples, both of whom are dead, have been collected and the information verified. This number constitutes 45 per cent of all the couples that it could have been at all possible to discover by the means employed.

Among the 2578 couples, the husband and wife of 7 died simultaneously or within 24 hours from homicide-suicide, accidents, etc. They have been excluded from further consideration and so the material on which this study is based is constituted by the death records of 2571 couples.

Correlation in the Age at Death of Husbands and Wives.—The product-moment coefficient of correlation between the length of life in husbands and in wives is found to equal $+0.5594 \pm 0.0092$. The tendency for marital partners to have the same duration of life is apparently independent of whether the husband and wife died of the same or of different causes. When only the couples who died from different causes are considered the correlation coefficient $r = +0.5485 \pm 0.0103$ which is obviously only slightly inferior to the value observed for all couples. The correlation coefficient calculated for the age at death of husbands and wives who died from the same cause is $+0.6454 \pm 0.0202$ which is higher than that observed for all couples. It would seem then that the correlation in length of life of husbands and wives is not due to whatever marital association there may be relative to the cause of death. On the other hand, when husbands and wives do die from the same condition they demonstrate also a greater tendency to the same life span.

Causes of Death in Husbands and Wives.—The cause of death stated on the death certificate has been coded according to the International List (1929 rev.); and when two or more causes were mentioned, the rules given in the third edition of the Manual of Joint Causes of Death (1933) have been followed.

The frequency with which the stated causes of death occurred in the 2571 married couples is shown in table 1.

In table 1 are given the observed number of husbands and wives who died from stated causes and the number expected assuming random matings according to cause of death. When the observed and the expected numbers are compared for all of the combinations in table 1 and the chi-square

TABLE 1

OBSERVED AND EXPECTED (IN PARENTHESIS) NUMBER OF HUSBANDS AND WIVES WHO DIED FROM STATED CAUSES

CAUSE OF DEATH OF HUSBAND	CAUSE OF DEATH OF WIFE: TUBERCULOSIS (ALL FORMS)	ARTERIO-SCLEROSIS	INFLUENZA AND PNEUMONIA (ALL FORMS)	CANCER AND OTHER MALIGNANT TUMORS	NEPHRITIS	CEREBRAL HEMORRHAGE	HEART DISEASES	OTHER AND ILL-DEFINED CAUSES	ALL CAUSES
Tuberculosis (all forms)	20 (7.4)	4 (4.5)	11 (10.1)	8 (12.5)	7 (10.6)	13 (14.9)	29 (31.0)	25 (26.0)	117
Arteriosclerosis	10 (8.9)	6 (5.3)	14 (12.8)	10 (14.9)	16 (12.7)	19 (17.9)	32 (37.1)	33 (31.2)	140
Influenza and pneumonia (all forms)	12 (11.4)	9 (6.9)	27 (15.5)	15 (19.2)	16 (16.3)	13 (23.9)	40 (47.7)	48 (40.1)	180
Cancer and other malignant tumors	9 (11.9)	5 (7.1)	10 (16.1)	29 (19.9)	14 (16.9)	25 (23.9)	64 (49.5)	31 (41.6)	187
Nephritis	15 (19.0)	12 (11.4)	30 (25.8)	25 (31.9)	33 (27.1)	36 (38.2)	77 (79.2)	71 (66.5)	299
Cerebral hemorrhage	17 (25.4)	14 (15.2)	23 (34.5)	50 (42.6)	46 (36.2)	56 (51.0)	108 (106.0)	86 (89.0)	400
Heart diseases	39 (38.4)	22 (23.4)	44 (53.0)	66 (65.5)	50 (55.6)	84 (78.4)	184 (162.6)	125 (136.6)	614
Other and ill-defined causes	41 (40.2)	26 (24.2)	63 (54.7)	71 (67.6)	51 (57.4)	82 (80.9)	147 (167.9)	153 (141.0)	634
All causes	163	98	222	274	233	328	681	572	2571

test applied, it is found that the deviation between the observed and the expected amounts to 4.5 times its standard error. Therefore, from a statistical standpoint, it appears that relative to cause of death the husbands and wives of this sample do not constitute a random combination of men and women.

A more exact evaluation of this apparent association between husbands and wives relative to cause of death may be obtained when the data regarding each of the seven specific causes are arranged in a four-fold tabulation and the chi-square is calculated for each four-fold table. It can then be shown that the association between husband and wife with respect to tuberculosis lies far beyond the limits of chance deviation ($P = <0.0001$). The frequency with which both husbands and wives have died of influenza and pneumonia is also significantly larger than expected on the basis of random assortment ($P = 0.0016$). With regard to the concurrence of deaths of husbands and wives from cancer and from heart diseases, respectively, P approximates 0.025 and the deviations from chance expectancy may also be regarded as statistically significant. Instead, for nephritis, cerebral hemorrhage and arteriosclerosis, respectively, the concurrence in husbands and wives does not deviate significantly from what would be expected assuming random mating. In sum, these data reveal that when a husband or wife died of tuberculosis the relative number of their spouses who also died from this cause is three times as high as that of the spouses of persons who died from other causes. Influenza and pneumonia as a cause of death occur almost twice more often among the husbands and wives of individuals who died from this cause than among the deceased spouses of those who did not. This ratio is 1.5 in the case of cancer and 1.2 in the case of heart diseases.

To determine whether age and the time and order of death of the two spouses could affect the frequency with which both died from the same cause, a supplementary control sample to be used in lieu of the calculated expected was formed according to the following procedure:

1. All the records concerning husbands whose death preceded that of the wives were alphabetized and the same was done for the records of wives whose death preceded that of the husbands.

2. For each husband whose death preceded that of the wife, a widow was sought in the files of unmatched widows (already alphabetized as it will be recalled) having the following characteristics: (*a*) the age at death within ± 5 years that of the wife, (*b*) time of death within ± 5 years of that of the wife. The first widow in alphabetical order who had these characteristics was selected as a control.

3. For each wife whose death preceded that of the husband a control widower was selected in the same manner as described in 2.

Among the control couples thus obtained, the number found in which

both the male and female died from (*a*) tuberculosis, (*b*) influenza and pneumonia, (*c*) cancer and (*d*) heart diseases are shown in table 2 and compared to the number of marital couples both of whom have died from one of these causes and to the number expected assuming random mating.

TABLE 2

NUMBER OF COUPLES IN WHICH THE MAN AND WOMAN BOTH DIED FROM THE SAME CAUSE. *A* = OBSERVED NUMBER OF HUSBANDS AND WIVES. *B* = EXPECTED NUMBER OF HUSBANDS AND WIVES ASSUMING RANDOM ASSORTMENT OF MEN AND WOMEN. *C* = NUMBER OF CONTROL, NON-MARITAL, COUPLES

CAUSE OF DEATH	*A*	*B*	*C*
Tuberculosis (all forms)	20	7.4	9
Influenza and pneumonia (all forms)	27	15.5	19
Cancer and other malignant tumors	29	19.9	19
Heart diseases	184	162.6	166

From table 2 it is immediately apparent that the number of control couples and the number of the expected are to all intents and purposes the same. The closeness in the agreement between the control and calculated data is remarkable and shows that what differences there may exist relative to the age and period of death of the spouses do not alter the deviations found between observed and expected, and it also indicates that within the limits of this comparison the calculated numbers actually represent the expectations assuming random heterosexual mating.

Discussion.—There are two series of findings resulting from this first analysis of the mortality in husbands and wives. The first concerns the correlation in the length of life of the two marital partners. This correlation is found to be positive and high and its value is only slightly increased when the cause of death is the same in the husband and wife, and it is very little decreased when the cause is different. Since the same specific causation or disease entity (so far as can be learned from the death certificate) is not apparently responsible for the correlated length of life in husbands and wives it would seem appropriate to have recourse to those vague but understandable terms such as resistance or vitality and say that the phenomenon observed indicates the segregation and pairing, through marriage, of individuals having a similar degree of vitality or resistance to fatal pathological processes.

The second series of findings brings out that besides the existence of a high correlation in the length of life of husbands and wives there is a tendency for marital partners to die from the same cause when one of the mates dies from either tuberculosis, influenza and pneumonia, cancer, or heart diseases. For the last three groups of causes such a relationship has never been observed before, at least to the writer's knowledge, although it could have been expected regarding influenza and pneumonia. For the three causes mentioned, particularly for heart diseases, the tendency is not very

marked but is apparently significant from a statistical standpoint. The concentration of tuberculosis deaths among marital partners has long been recognized and students of tuberculosis have always been surprised to find that this association is not higher.

The hypotheses that on the basis of information now available could reasonably be formulated to elucidate the findings of this study may be summed up as follows:

1. The correlation in length of life or general vitality of husbands and wives could be due to conscious or unconscious marital selection although the economic level or economic changes undoubtedly play some part.

2. Tuberculosis is a chronic infectious disease and the association between husbands and wives relative to the mortality from this cause could be thus easily explained. But, in addition, marital selection both from the somatic and social aspects are probably important.

3. In the case of influenza and pneumonia the transmission of the pathogenic agents from one spouse to the other could be assumed to be the factor involved in the association noted. The transmission, however, due to the interval observed between the deaths of the two partners must be indirect; unless it is believed that one partner becomes a carrier. On the other hand, the common environment and the general economic level play a rôle in the incidence of this disease as apparently does also the somatic constitution. It is of interest to point out that the highest association for cause of death in husbands and wives is found relative to tuberculosis and influenza and pneumonia. Both are infectious diseases and for both one can postulate a number of factors that could produce the concurrence.

4. With respect to heart diseases, marital contagion might be an element if the etiological factor in rheumatic heart disease is definitely shown to be infectious, otherwise the somatic and constitutional aspects of marital selection might be considered.

5. None of the factors mentioned appear to explain the association between husbands and wives relative to death from cancer. Yet on logical grounds it would seem that the hypotheses outlined fairly exhaust the possibilities in this respect and consequently one of them must provide the necessary answer.

The complete inability to give an explanation concerning the cancer findings on the basis of present knowledge and the uncertainty regarding the views advanced about the other causes, heart diseases especially, while immediately discouraging, also seem to point the way for further investigations on these as well as other diseases. In particular, these observations serve to delineate more sharply the practical importance of studying the familial aggregate in relation to problems of health and disease.

* A full report will appear in a forthcoming number of *Human Biology*.

ON AN INEQUALITY OF GRUNSKY

BY D. C. SPENCER

DEPARTMENT OF MATHEMATICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Communicated September 6, 1940

1. Suppose that W is an arbitrary Riemann domain. We let $n(w)$ be the number of times W covers the point w, and define*

$$p(R) = \frac{1}{2\pi}\int_{-\pi}^{\pi} n(Re^{i\Phi})d\Phi. \tag{1.1}$$

For simplicity we write $n_0 = n(0)$. The purpose of this brief note is to prove the following theorem:

THEOREM 1. *Suppose that $n_0 > 0$, and that*

$$\int_0^{R_1} p(R)d(\pi R^2) \leqslant n_0\pi R_1^2 \tag{H}$$

for all $R_1 > 0$ (mean n_0 − valency). Then

$$\exp\left\{\frac{1}{n_0}\int_0^{\infty} \lg R^2 d[-p(R)]\right\} \leqslant \frac{1}{n_0}\int_0^{\infty} R^2 d[-p(R)], \tag{1.2}$$

and there is equality if, and only if, the left-hand side is infinite, or $p(R) \equiv n_0$ for $0 \leqslant R < R_0 < \infty$, and $p(R) \equiv 0$ elsewhere.†

Since

$$\frac{1}{n_0}\int_0^{\infty} d[-p(R)] = \left[-\frac{1}{n_0}p(R)\right]_0^{\infty} = 1,$$

Theorem 1 is a case of the theorem of the arithmetic and geometric means with weight function $d[-p(R)]$ which, it is easy to see, is not necessarily of constant sign. The inequality (1.2) (with a different formulation) has been proved by Grunsky‡ subject to the stronger hypothesis that $n(w) \leqslant n_0 = n(0)$. My main purpose is to give an extremely simple proof of this result; the generalization (to mean n_0-valency) follows incidentally. If we write $w = Re^{i\Phi}$, and denote by B the boundary of W, the inequality (1.2) can be written in the alternative form:

* Compare Spencer, reference 4 at end of paper.

† "$f \equiv g$" here means $f = g$ except in a null set. If we add to Theorem 1 the hypothesis that $p(R) \equiv n_0$ in a neighborhood of $R = 0$, we may state it as a theorem in real variables without reference to Riemann domains. It is possible to develop the theorem further in this direction.

‡ See reference 3 at end of paper.

$$\exp\left\{\frac{1}{2n_0\pi}\int_B \lg R^2 d\Phi\right\} \leqslant \frac{1}{2n_0\pi}\int_B R^2 d\Phi,$$

where (for example) we define $\int_B = \lim \int_{B'}$ as $B' \to B$ from the interior of W. We may also interpret $\int_B \ldots d[-p(R)]$ as $\lim \int_{B'}$.

2. We may suppose that

$$\frac{1}{n_0}\int_0^\infty R^2 d[-p(R)] = R_0^2 < \infty; \tag{2.1}$$

otherwise the theorem is true (and trivial). We then define

$$p_0(R) = \begin{cases} n_0, \text{ if } 0 \leqslant R < R_0, \\ 0, \text{ otherwise.} \end{cases}$$

Since W covers $w = 0$ with multiplicity n_0, there exists a closed circle of radius r and center $w = 0$ every point P of which is covered by W at least n_0 times. Since, by (H), it is covered on the average at most n_0 times, it follows that the circle is covered exactly n_0 times. Hence $p(R) = n_0$ for $0 \leqslant R \leqslant r$, and so

$$\int_0^\infty \lg R^2 d[-p(R)] = \int_r^\infty \lg R^2 d[-p(R)] = [\lg R^2[-p(R)]]_r^\infty + \int_r^\infty p(R) d[\lg R^2]$$

by a partial integration;

$$= \int_r^\infty p(R) d[\lg R^2] - n_0 \lg \frac{1}{r^2}, \tag{2.2}$$

since $p(\infty) = 0$, and $p(r) = n_0$. (2.2) may be written in the form

$$\int_r^\infty p(R)\frac{d(R^2)}{R^2} - n_0 \lg \frac{1}{r^2},$$

and a second partial integration now gives

$$\left[\frac{1}{R_1^2}\int_r^{R_1} p(R)d(R^2)\right]_{R_1=r}^{R_1=\infty} + 2\int_r^\infty \left\{\int_r^{R_1} p(R)d(R^2)\right\}\frac{dR_1}{R_1^3} - n_0 \lg \frac{1}{r^2} = 2\int_r^\infty \left\{\int_r^{R_1} p(R)d(R^2)\right\}\frac{dR_1}{R_1^3} - n_0 \lg \frac{1}{r^2}. \tag{2.3}$$

Next,

$$\int_r^{R_1} p(R)d(R^2) = \int_0^{R_1} p(R)d(R^2) - \int_0^r p(R)d(R^2) = \int_0^{R_1} p(R)d(R^2) -$$

$$n_0 r^2 \leqslant \text{Min}\left\{n_0 R_1^2, n_0 R_0^2\right\} - n_0 r^2 = \int_r^{R_1} p_0(R)d(R^2) \quad (2.4)$$

by (H) and (2.1), since

$$\int_0^{R_1} p(R)d(R^2) \leqslant \int_0^{\infty} p(R)d(R^2) = \int^{\infty} R^2 d[-p(R)] = n_0 R_0^2.$$

Substituting from (2.4) into (2.3), we obtain that

$$\int_0^{\infty} \lg R^2 d[-p(R)] \leqslant \int_0^{\infty} \lg R^2 d[-p_0(R)], \quad (2.5)$$

and therefore

$$\exp\left\{\frac{1}{n_0}\int_0^{\infty} \lg R^2 d[-p(R)]\right\} \leqslant \exp\left\{\frac{1}{n_0}\int_0^{\infty} \lg R^2 d[-p_0(R)]\right\} =$$

$$\exp\left\{\lg R_0^2\right\} = R_0^2 = \frac{1}{n_0}\int_0^{\infty} R^2 d[-p(R)]$$

by (2.1). This proves the theorem, apart from the statement about the case of equality.

But this is easily disposed of. For if $p(R) \not\equiv p_0(R)$, then for a set E of R_1 of positive measure

$$\int_0^{R_1} p(R)d(R^2) < n_0 R_1^2, \quad (2.6)$$

if $R_1 \epsilon E$. Since (2.6) plainly implies inequality at (2.5), it therefore implies it at (1.2).

3. I conclude with an application of Theorem 1. Suppose that $f(z)$, regular in $|z| < 1$, maps the unit circle $|z| < 1$ on a (Riemann) domain W, and that $f(0) = 0$. Let W^* be the domain composed of those points of W which are visible to an eye placed at $w = 0$ if all boundary elements B of W are opaque, and let $Z^* = f^{-1}(W^*)$. Then Z^* and W^* are schlicht domains which contain the points $z = 0$ and $w = 0$, respectively, and W^* is the "star" of W with respect to the point $w = 0$. If we write $V(W^*)$ and $V(Z^*)$ for the areas of W^* and Z^*, then we have the following result (a generalization of the lemma of Schwarz):

THEOREM 2.*

$$|f'(0)| < \sqrt{\frac{V(Z^*)}{\pi}} \cdot \sqrt{\frac{V(W^*)}{\pi}}, \tag{3.1}$$

unless $f(z) = a_1 z$.

We write D^* for the "star" of a domain D with respect to the origin,† $\overline{D}$ for the closure of D (i.e., $\overline{D} = D +$ boundary of D), and write D^{-1} for the set of points $R^{-1}e^{i\Phi}$, where $Re^{i\Phi} \epsilon D$. We denote the z-plane, including the point $z = \infty$, by P_z, and define

$$Z^{**} = P_z - \overline{\left\{[(P_z - \overline{Z}^*)^{-1}]^*\right\}^{-1}}.$$

Then $Z^* \subset Z^{**}$. Bermant‡ has proved that

$$|f'(0)| \leqslant \sqrt{\frac{V(Z^{**})}{\pi}} \cdot \sqrt{\frac{V(W^*)}{\pi}}, \tag{3.2}$$

with equality only for $f = a_1 z$.

To prove Theorem 2, we need only apply Theorem 1 at the appropriate point in Bermant's proof of (3.2) (Bermant, *loc. cit.*). However, for the sake of completeness, I include a proof of Theorem 2 using a method§ different from that of Bermant.

We may plainly suppose that $|f'(0)| = 1$. We write $\zeta = \xi + i\eta$. Suppose that a function $g(\zeta)$ maps a Riemann domain G_1 on a (Riemann) domain G_2. Let $q(\eta)$ be a set of points of G_1, defined for each η of a set H, which lie over the line $\eta =$ constant, and write

$$E_1 = \sum q(\eta),\ V_1 = mE_1,\ V_2 = \int\int_{E_1} |g'(\zeta)|^2 d\xi d\eta$$

$$l(\eta) = g\{q(\eta)\},\ L(\eta) = ml(\eta) = \text{"length" of } l(\eta).$$

Then we have:

* Theorem 2 is stated without proof in my paper (reference 5).

† But Z^* is *not* a star; this will lead to no confusion.

‡ Bermant (see reference 1 at end of paper). The first result of this kind is due to Golusin (see reference 2), who proved that $|f'(0)| \leqslant \sqrt{\frac{V(W^*)}{\pi}}$.

§ See Spencer (reference 5), where the method is applied to multiply-connected domains.

LEMMA. *Suppose that* $L(\eta) \geqslant \lambda$ *for almost all* $\eta \in H$. *Then*

$$(\lambda \cdot mH)^2 \leqslant V_1 V_2$$

(*and there is equality only if* $V_1 = 0$ *or* $g = a_0 + a_1\zeta$).

I omit the proof (which is two lines in length); it has been given elsewhere* in any case.

We suppose that r is small enough that the closed circle $|w| \leqslant r$ is contained in W^*. We remove this circle from W^*, and cut the remaining domain along the real axis from $w = r$ to the nearest boundary point to form a simply connected domain W^*_r. We write $Z^*_r = f^{-1}(W^*_r)$, transform W^*_r, Z^*_r by $\zeta = \lg w$, $\sigma = \lg z$ into two domains $G_1(r)$, $G_2(r)$, respectively, and apply the lemma. We take for H the interval $(0, 2\pi)$, for $q(\eta)$ those points of $G_1(r)$ which lie on the line $\eta = \text{constant}$. Then $q(\eta)$ is (for every η of H) a line segment connecting the logarithmic transform of some *boundary point* of W to the straight line $\zeta = \lg r$. Therefore the transform $l(\eta)$ of $q(\eta)$ connects the line $Re(\sigma) = 0$ to a curve lying inside the rectangle

$$\lg r - \epsilon(r) \leqslant Re(\sigma) \leqslant \lg r + \epsilon(r),\ 0 \leqslant Im(\sigma) \leqslant 2\pi,$$

where $0 < \epsilon(K \cdot r) < 0(r)$, and so $L(\eta) \geqslant \lg \frac{1}{r} - \epsilon(r) = \lambda$. By the lemma we have, therefore,

$$2\pi\left(\lg \frac{1}{r} - \epsilon(r)\right) \leqslant \sqrt{V_1(r) \cdot V_2(r)} \leqslant \frac{V_1(r) + V_2(r)}{2}, \qquad (3.3)$$

where $V_1(r)$ and $V_2(r)$ are the areas of $G_1(r)$ and $G_2(r)$, respectively. Let $p_1^*(R)$, $p_2^*(R)$ be the functions defined by (1.1) in terms of W^*, Z^*, respectively. Then

$$V_1(r) = \int_r^\infty p_1^*(R) d(2\pi \lg R) = 2\pi \lg \frac{1}{r} + \pi \int_r^\infty \lg R^2 d[-p_1^*(R)], \qquad (3.4)$$

and

$$V_2(r) = \int_r^\infty p_2^*(R) d(2\pi \lg R) + 0(r) = 2\pi \lg \frac{1}{r} + \pi \int_r^\infty \lg R^2 d[-p_2^*(R)] + 0(r). \qquad (3.5)$$

* Spencer (reference 5).

Substituting from (3.4) and (3.5) into (3.3), and letting $r \to 0$, we obtain that

$$0 \leqslant \int_r^\infty \lg R^2 d[-p_1^*(R)] + \int_r^\infty \lg R^2 d[-p_2^*(R)]. \qquad (3.6)$$

Since both W^* and Z^* satisfy the hypothesis (H)† with $n_0 = 1$, we may take the exponent of both sides of (3.6) and apply Theorem 1, obtaining (3.1). There is equality (by Theorem 1) only if W^* is a circle—i.e., only if $f = a_1 z$. This completes the proof of Theorem 2.

Inequality (3.6) (in different notation) is due to Bermant (loc. cit.). We note that $d[-p_1^*(R)] \geqslant 0$, but the corresponding statement for $p_2^*(R)$ is in general false. It is for this reason that the ordinary version of the theorem of the arithmetic and geometric means (in which the weighting is assumed positive) cannot be applied.

[1] Bermant, A., "Remarque sur le lemme de Schwarz," *C. R. Acad. Sci. Paris*, **207**, 31–33 (1938).

[2] Golusin, G. M., "Einige Überdeckungssätze für die im Kreise regulären Funktionen," *C. R. Acad. Sci.* U. R. S. S., **2**, 617–619 (1937).

[3] Grunsky, H., "Neue Abschätzungen zur konformen Abbildung ein- und mehrfach zusammenhängender Bereiche," *Schr. math. Sem. Univ. Berlin*, **1**, 95–140 (1932).

[4] Spencer, D. C., "On finitely mean valent functions," *Proc. London Math. Soc.* (to appear shortly).

[5] Spencer, D. C., "On a Theorem of Rengels," *Journal of Math. and Physics, Mass. Inst. of Technology* (to appear shortly).

† In fact, both satisfy the stronger hypothesis that $n(w) \leqslant 1$; we require, therefore, only Grunsky's theorem.

MINIMAL CROSS-CUT SUBGROUPS RELATIVE TO THE PRODUCT OF THEIR ORDERS

BY G. A. MILLER

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ILLINOIS

Communicated August 23, 1940

If H_1 is a subgroup of a given group G then according to E. Galois (1811–1832) all the operators of G can be represented in the form of right co-sets or in the form of left co-sets with respect to H_1. This is equivalent to the earlier arrangement by P. Abbati (1768–1842) of all the operators of G in the form of a rectangle either with respect to right or left multiplication in which the operators of H_1 constitute the first row and always appear on the same side in the products. Let H_2 be any other subgroup of G which is not

contained in H_1. The operators of H_2 will then be equally distributed in the given rectangular arrangement of the operators of G among the rows in which at least one operator of H_2 appears. That is, each of these rows contains either no operator of H_2 or it contains just as many operators of H_2 as appear in the first row.

When H_2 coincides with G then the operators of H_2 appear in all of the rows of G and this may also be the case when H_2 is a proper subgroup of G, that is, when H_2 is neither the identity nor the entire group. In what follows the term subgroup will always be used for a proper subgroup unless the contrary is explicitly stated. A necessary and sufficient condition that the operators of H_2 appear in all of the rows of G in the given arrangement is that the product of the orders of H_1 and H_2 divided by the order of their cross-cut is equal to the order of G. When this condition is satisfied H_1 and H_2 may be said to have a minimal cross-cut with respect to the product of their orders or to be minimal cross-cut subgroups relative to the product of their orders. It is obvious that the product of the orders of two subgroups divided by the order of their cross-cut cannot exceed the order of the group. It is a multiple of the order of each of these subgroups and hence composite.

When G is the non-cyclic group of order p^2, p being a prime number, or when it is the cyclic group whose order is the product of two distinct prime numbers then every pair of subgroups of G is a pair of minimal cross-cut subgroups relative to the product of their orders. It is not difficult to prove that no other group has this property. It is obvious that the order of such a group could not be of the form p^m, $m > 2$, since it could not be cyclic because one of the two subgroups could not appear in the other as H_2 was supposed to be not contained in H_1. It could therefore not be the group of order 2^m which contains only one subgroup of order 2. Since every other non-cyclic group of order p^m contains the non-cyclic group of order p^2 it has been proved that the order of G could not be p^m, $m > 2$. As the order of G could not be divisible by three distinct prime numbers or by the square of a prime number when it is divisible by just two distinct prime numbers it has been proved that *the non-cyclic group whose order is the square of a prime number and the cyclic group whose order is the product of two distinct prime numbers are the only groups which have the property that every pair of their subgroups is a pair of minimal cross-cut subgroups with respect to the product of the orders of these subgroups.*

When operators of H_2 appear in all of the rows of G with respect to H_1 the operators of H_1 appear in all of the rows of G with respect to H_2 so that the property of minimal cross-cut subgroups is a reciprocal property. If one of a pair of minimal cross-cut subgroups relative to the product of their orders is of prime order then the other is of the same prime index. It could not be of a larger index since its order is a divisor of the order of the group,

and it could not be of a smaller index since the product of the orders of the two subgroups divided by the order of their cross-cut is equal to the order of the group. In particular, if one of two minimal cross-cut subgroups relative to the product of their orders is of order 2 the other is an invariant subgroup of index 2, and if one of two minimal cross-cut subgroups with respect to the product of their orders is of index 2 then every other subgroup of the group forms with this subgroup a pair of minimal cross-cut subgroups. In particular, the minimal cross-cut subgroups with respect to a given subgroup may have different orders and some other subgroups may have a smaller cross-cut.

Suppose that H_1 and H_2 are two conjugate subgroups of G. It is then easy to prove that H_1 and H_2 cannot have a minimal cross-cut relative to the product of their orders. If they had such a cross-cut the operators of H_2 would be equally distributed among all the rows of G with respect to the rectangle in which H_1 is the first row and the operators of H_1 appear on the left in the products which constitute the other rows. All the operators of each of these rows would transform H_1 into the same one of its conjugates. These could not include H_2 since each row may be assumed to begin with an operator of H_2 and hence cannot transform H_1 into H_2 since it transforms H_2 into itself. It therefore results that the cross-cut of any two conjugate subgroups of G is larger with respect to the product of their orders than the index of one of these subgroups under the entire group. That is, the quotient of this product of these orders and the order of this cross-cut is less than the index of one of these conjugate subgroups under the entire group.

If G is a multiply transitive permutation group of degree n and H_1, H_2 are two subgroups composed of all its permutations which omit a given letter in each case then H_1 and H_2 are known to be conjugate and if the permutations of H_1 constitute the first row when G is written in the ordinary rectangular form then the permutations of H_2 will appear in $n - 1$ of the rows, since H_1 and H_2 are transitive. Moreover, if G is not multiply transitive then the permutations of H_2 cannot appear in as many as $n - 1$ of these rows because the number of rows in which these permutations appear is equal to the degree of a transitive constituent if H_1 and this number cannot exceed $n - 3$. The property of multiple transitivity as well as the degree of the various transitive constituents of H_2 can therefore be directly observed from the number of rows in which the permutations of H_2 appear with respect to H_1.

A cyclic prime power group does not contain a pair of minimal cross-cut subgroups relative to the product of their orders since one of every two subgroups of such a group is contained in the other. Every other abelian group contains at least one pair of such subgroups since it is the direct product of two cyclic groups which have only the identity in common. Moreover, every non-cyclic prime power group contains a pair of minimal

cross-cut subgroups relative to the product of their orders since such a group of order p^m contains two distinct subgroups of index p which can be used for H_1 and H_2, respectively. Since a solvable group necessarily contains an invariant subgroup of prime index this subgroup and an operator which is a power of this prime number but is not contained in this subgroup can be used as a pair of minimal cross subgroup relative to the product of their orders whenever this solvable group is not a cyclic group of prime power order.

The symmetric group of degree $n > 3$ contains a regular group which is one of a pair of minimal cross-cut subgroups with respect to the product of their orders if the other is the symmetric group of degree $n - 1$. This is also true of the alternating group of degree n whenever n is odd. When n is even and equal to $2m$ we can construct a group of order $2^m m!$ and of degree n which contains a subgroup of index 2 composed of its positive permutations which with the alternating group of degree $n - 1$ constitutes a pair of minimal cross-cut subgroups with respect to the product of their orders. These two subgroups have in common a group of order $2^{n-1}m!$ divided by n and hence the product of their orders is $n!$ over 2. Therefore every symmetric group and every alternating group contains a pair of minimal cross-cut subgroups relative to the product of the orders of these subgroups.

One of the most important properties of two minimal cross-cut subgroups relative to the product of their orders is that they cannot be conjugate under the group. If one of these two subgroups and their cross-cut are invariant under the group the other subgroup is isomorphic with the quotient group with respect to this invariant subgroup. If both of these minimal cross-cut subgroups are invariant under the group the commutators which have one element from each of these subgroups are in their cross-cut and this cross-cut is also invariant under the group. If this cross-cut is the identity then every operator of one of these subgroups is commutative with every operator of the other and hence the group is the direct product of these two minimal cross-cut subgroups relative to the product of their orders. For instance, a group whose order is divisible by two and only two distinct prime numbers which has invariant Sylow subgroup of different orders is the direct product of two such subgroups.

The difference between the product of the orders of two subgroups and the order of the group is divisible by the order of each of these subgroups. When this difference is zero and the orders of these subgroups are relatively prime they are minimal cross-cut subgroups relative to the product of their orders. The product of the orders of two subgroups of a group cannot exceed the square of the order of the group divided by 4 and when it is equal to this number the group contains an invariant subgroup of index 4 corresponding to the non-cyclic group of order 4 as a quotient group. In par-

ticular, the group contains then at least three subgroups of index 2 and the two given subgroups are minimal cross-cut subgroups relative to the product of their orders. *The square of the order of a group divided by the product of the orders of two subgroups is a composite number and when this number is divisible by only two prime numbers (equal or distinct) these subgroups are necessarily minimal cross-cut subgroups relative to the product of their orders.*